MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ARTIFICIAL INTELLIGENCE LABORATORY

A. I. Memo No. 597 October, 1980

REPRESENTATION AND RECOGNITION OF THE MOVEMENTS OF SHAPES

DAVID MARR AND LUCIA VAINA

The problems posed by the representation and recognition of the movements of 3-D shapes are analyzed. A representation is proposed for the movements of shapes that lie within the scope of Marr & Nishihara's (1978) 3-D model representation of static shapes. The basic problem is, how to segment a stream of movement into pieces each of which can be described separately. The representation proposed here is based upon segmenting a movement at moments when a component axis, e.g., an arm, starts to move relative to its local coordinate frame (here, the torso). So that for example walking is divided into a sequence of the stationary states between each swing of the arms and legs, and the actual motions between the stationary points (relative to the torso, not the ground). This representation is called the state-motion-state (SMS) moving shape representation, and several examples of its application are given.

This report describes research done at the Artificial Intelligence laboratory of the Massachusettes Institute of Technology. Support for this work was provided in part by the Office of Naval Research under ONR Contract N00014-72-C-0260 and the National Science Foundation under NSF Grant 79231-10-MCS. Support for the Laboratory's artificial intelligence research is also provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research Contract N00014-75-C-0643.

Introduction

An important aspect of vision is its ability to inform us of the shapes and spatial arrangements of physical objects. In addition, we can notice if an object moves, and we can see how. The fact that we can perceive and recognize three-dimensional shapes means two things; firstly, there must be a symbolic system for *representing* the shape information in the brain, and secondly the brain must contain a set of processes capable of *deriving* this information from images.

In their study of how to represent three-dimensional shape information, Marr and Nishihara (1978) laid down three criteria that such representations should satisfy in order to account for the efficiency with which the human visual system recognises 3-D objects:

Criterion 1 (Accessibility). The representation should be easy to compute from the pictorial image.

Criterion 2 (Scope and uniqueness). It should provide a description of a sufficiently large class of shapes, and for each shape within its scope, the representation should provide a description that is unique from *any point of view*. Otherwise, if the description is to be used for recognition, the difficult problem will at some point arise of whether two descriptions describe the same shape.

Criterion 3 (Stability and sensitivity). The representation should reflect the similarity between two like shapes whilst also preserving the differences. As Sutherland (1979) put it, it is important to be able to recognize both that a shape is a man and that it is Jones or Smith.

In the light of these criteria, Marr and Nishihara considered three aspects of a representation's design; (i) the representation's coordinate system, (ii) its primitives, which are the primary units of shape information used in the representation, and (iii) the organization the representation imposes on the information it describes. They concluded that for recognition, a shape representation should be based on an object-centred rather than a viewer-centred coordinate system, that it should include volumetric primitives, not just the type of surface primitive more easily derivable from images, and that it should impose a modular heirarchical organization on the description. These aspects of a shape representation are captured in almost their simplest form by the *3-D model representation*, illustrated in figure 1.

The basic unit of this representation is the *3-D model*, which consists of two parts. Firstly, an overall *model axis*, shown on the left of each box in figure 1, attached to which is a rough volumetric primitive (the cylinder) describing coarsely the size and orientation of the overall shape represented. Secondly, a collection of *component* axes, as shown on the right of each box, which give more detailed information about the spatial organization of the shape. Each component axis is also attached to a volumetric primitive (a cylinder here), and its location in space is defined relative to the principal axis of the model. The principal axis is the axis which has the most adjoining axes, so for the human 3-D model, it would be the torso axis.

Much of Marr and Nishihara's article is concerned with how this representation satisfies the criteria they laid down. Roughly, the scope of the representation is restricted to shapes that have a natural or canonical axis--as defined, for example, by elongation or symmetry, or even the gravitational vertical. Uniqueness is achieved largely because the representation is object-centred. The trade-off between stability and sensitivity is accomplished by looking at different levels in the representation; to see whether the shape is a man, one looks at the topmost level; to decide whether he is a bricklayer or a concert pianist, one looks at the 3-D model of his hands. And finally, although the accessibility issue has not been fully resolved, a start has been made on the problem of how to derive a shape's natural axis from an image (Marr, 1977), and there seems every reason

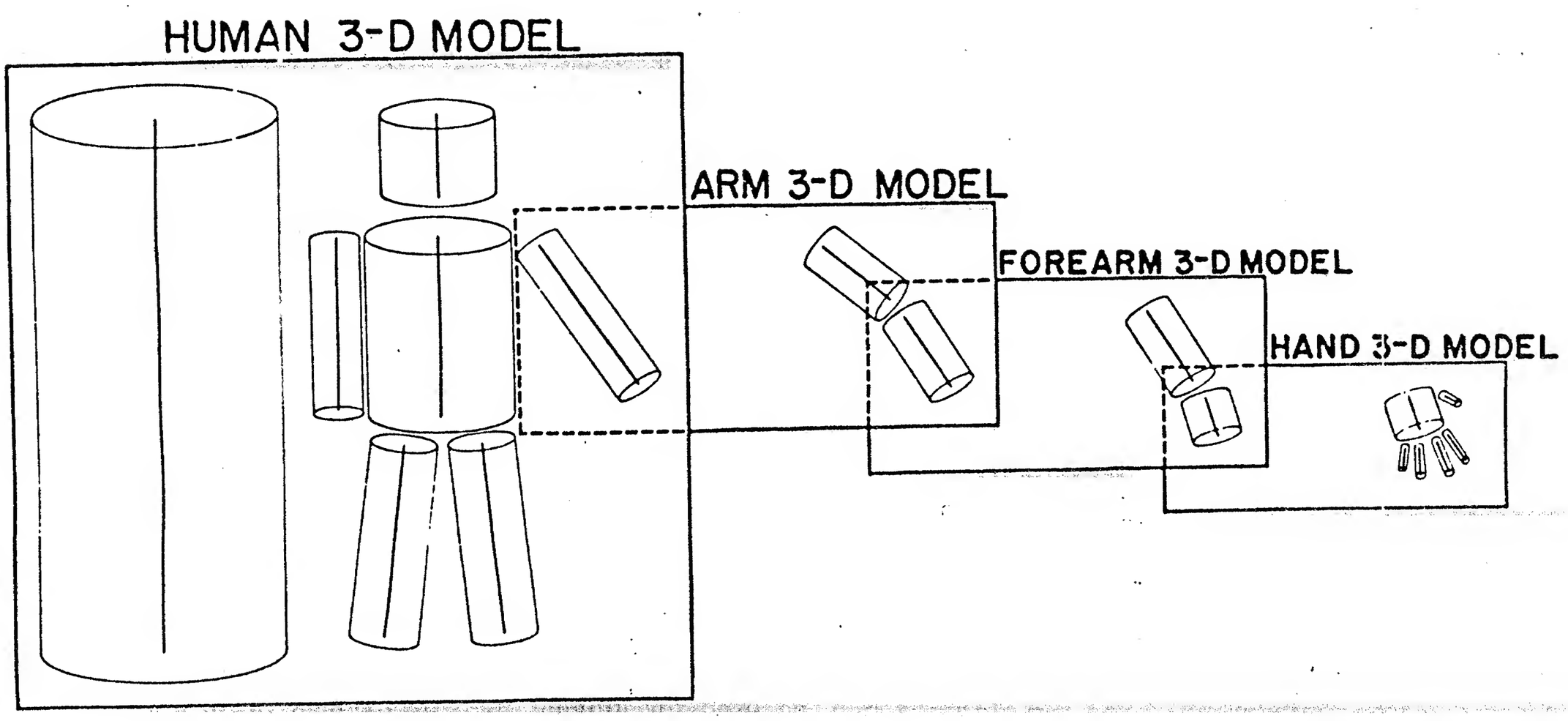


Figure 1. This diagram, taken from Marr & Nishihara, figure 3, illustrates the organization of shape information in a 3-D model description. Each box corresponds to a 3-D model; with its model axis on the left side of the box and the arrangement of its component axes are shown on the right side. In addition, some component axes have 3-D models associated with them and this is indicated by the ways the boxes overlap. The relative arrangement of each model's component axes, however, is shown improperly, since it should be in an object-centred system rather than the viewer-centred projected used here (a more correct 3-D model is shown in figure 2). The important characteristics of this type of organization are: (i) each 3-D model is a self-contained unit of shape information and has a limited complexity, (ii) information appears in shape contexts appropriate for recognition (the disposition of a finger is most stable when specified relative to the hand that contains it), and (iii) the representation can be manipulated flexibly. The approach limits the representation's scope however, since it will only be useful for shapes that have well-defined 3-D model decompositions.

to hope that as we expand our knowledge of how to derive shape information from images, the difficult problems posed by arbitrary vantage points will eventually yield to analysis.

Introducing Movement

Although the general topic of visual motion has been extensively studied in the past, interest has for the most part been confined to the problem of deriving surface shape information from a sequence of images, using measurements of the changing appearance of an object in motion (Ullman, 1979; Marr & Ullman, 1980; Longuet-Higgins & Prazdny, 1980). Our concern in this article is, however, with the representation of moving shapes. It arises because although we can perceive and recognize shapes, we are just as capable of perceiving and recognizing movements (e.g., actions or gestures).

As we have seen, Marr and Nishihara analyzed the problems posed by the representation of static three dimensional shapes. The representation scheme that they proposed forms the point of departure of our investigation, which is divided into two parts. In the first, we analyze the problems associated with the representation of instantaneously moving shapes; and in the second part we discuss how to combine descriptions of static and moving shapes into larger units, thus providing a primitive representation of movements that are extended in time. The ability to recognize and represent such movements introduces a host of new questions bearing on the more semantic and less purely visual aspects of objects and actions. In a subsequent article, we shall address the rather deep and fascinating issues to which this leads.

Representing the Instantaneous Motion of Shapes

The issues involved in representing moving shapes for recognition are, as one might expect, quite similar to those posed by static shapes. The same criteria may be used for judging the effectiveness of such a representation as the ones that we listed earlier in the article, and the design decisions that have to be made are similar to those involved in defining a representation for shapes.

The underlying representation of shape in the two cases will therefore be approximately the same, except in so far as its movement introduces additional or even a different decomposition of a shape into its components. In fact, the effects of motion are already incorporated to some extent in the 3-D model representation; the division of the arm at the elbow, which is visible in figure 1, occurs basically because the elbow is a joint and the forearm is hinged about this point.

For many shapes, therefore, the subdivision provided by their decomposition into generalized cylinder components will be roughly the same as those forced by the shape's articulation, and so our first task is to examine ways of adding the representation of movement to a 3-D model. We formulate the issues in the same way as did Marr & Nishihara (1978).

[1] Coordinate System

The natural coordinate system in the context of a 3-D model is an object-centred one. It is natural to describe the swinging of an arm component as back-and-forth relative to the torso axis, and the forearm as bending at the elbow relative to the upper arm. A representation system that is designed to capture canonically the intrinsic movements of a shape--whether a man is walking, running, jumping or limping, whether a swaying tree is liable to fall, the intention movements of a hunting animal--is virtually forced to use an object-centred system. The basic reason is similiar to but stronger than the reasons that apply to the representation of static

shapes. As Marr and Nishihara pointed out, in order to discriminate a left from a right hand using only a viewer-centred representation, many different views would have to be stored, whereas it is relatively straightforward in an object-centred representation. For moving shapes, ideas like "turn left" or "turn right" are inherently object-centred, and hard to represent in viewer-centred systems. And in addition to this, motions and gestures made by an animal are organized by the animal in its own coordinate system. A representation designed to facilitate the interpretation of such movements must perforce use one too.

Motion does however raise additional issues in a way that perhaps pure shape does not. Consider for example, the 3-D model for a whole man. If the man is still, then although his position relative to the viewer may be of interest, it is not something one would ordinarily include in a shape representation. If the man is walking, however, his speed and direction probably should be included, represented as the speed and direction of the model axis of the whole human shape. And it is not hard to think of situations where it is important to know this type of information relative to the viewer--is a tiger approaching or receding, for instance, or is the man's arm moving towards one as he hurls a spear.

We may conclude from this that, although the main need in representing movement is for an object-centred coordinate system in the real world, there will also arise the need for some viewer-centred information, usually about the overall motion of the whole shape.

[2] Primitives

Our task now is to design primitives for describing the instantaneous motion of a shape. If one thinks for a moment of a moving, deforming lump of putty, one can see that the general form of this problem is as intractable as the problem of representing arbitrary stationary shapes. In the spirit of simplicity, therefore, we shall limit the repertoire of shapes we consider to those of the 3-D model representation. This divides our problem into two halves, changes in the shape primitives themselves, and more interestingly, changes in their spatial organization.

There is one other rather general limitation that we shall impose on the representation. In general, the full specification of how a variable $x(t)$ is instantaneously changing with time involves knowing all its time derivatives - velocity, acceleration, and higher order terms. We shall on the whole limit ourselves to just velocities, partly for the sake of simplicity and partly because of accessibility issues. Human beings, for example, are rather poor at the visual estimate of acceleration. The only exceptions to this will be when we consider rotational movements; but although they do involve accelerations, they can equally well be described in terms simply of angular velocities.

Changes in shape primitives

If our representation is to deal adequately with changing shapes, it must be capable of describing the types of shape change involved in expanding a telescope, stretching a piece of elastic, inflating a balloon, bending a stick, squashing a ball of dough, and so forth. The shape primitives of the 3-D model representation were limited to generalized cones, which are the shapes swept out by moving a cross-section of constant shape but smoothly varying size along an axis. In practise, Marr and Nishihara restricted their analysis to specifying the rough length and width of a shape (hence the cylinders in figure 1), although it is clear that this can be extended. One more variable could for example serve to define a pillow-shaped region of space rather than a cylinder, or to introduce some curvature into the axis. In an interesting study of the shapes of ancient Greek pottery, Hollerbach (1975) constructed a set of shape primitives that was complex enough to provide a basis for the distinctions that archaeologists make between the various styles.

In the case where only one basic shape primitive is allowed, like the cylinder or pillow-shaped primitive, its size and shape are defined by just a few numbers. For the cylindrical primitive, Marr and Nishihara used a measure of overall size *s*, length *l*, and radius *r*, and for the pillow-shaped region, three local variables would be used, *l*, *r* and *w* say. In this simple situation, changes of shape can be described by the time derivatives *dl/dt*, *dr/dt* and *dw/dt*, but it is probably better to normalize them in some way, for example by using *l/l*, *r/r* and *w/w*. In these terms, expanding a telescope or stretching some elastic would yield *l/l* positive, and *r/r* approximately zero; inflating a balloon would yield *l/l*, *r/r* and *w/w* all positive and equal in value; and in squashing a ball of dough, *l/l* would be negative, whilst *r/r* and *w/w* would be positive.

The problems raised by bending a stick (i.e., curving the axis of a shape component of the representation) depend on how one parametrises the curved axis, and on how many primitives are used to represent different types of curvature. If only one parametrized primitive is used, as we discussed above for the cylinders, then the introduction of changes in the parameter values poses no new problems. If one has a variety of parametrized primitives however, as Hollerbach did, then changes in shape can induce shifts from one shape primitive to another. We shall not consider this type of complication further here.

Finally, there is the question of accessibility. If the parameters associated with a shape primitive are accessible at all from one or more images, then in principle so are their time derivatives. We hold no illusions, however, that these numbers will often be available accurately. Whether they are positive or negative, or small, medium or large is often all one can hope to obtain, and indeed probably all one needs (see later). Techniques based for example on directional selectivity can yield quite powerful determinants of the signs of such quantities (Marr & Ullman, 1980); and the variable precision which other methods can supply can be accommodated by a system for representing parameter values that includes tolerances, in the manner of figure 2 (Marr and Nishihara's figure 5).

[3] Changes in Spatial Organization

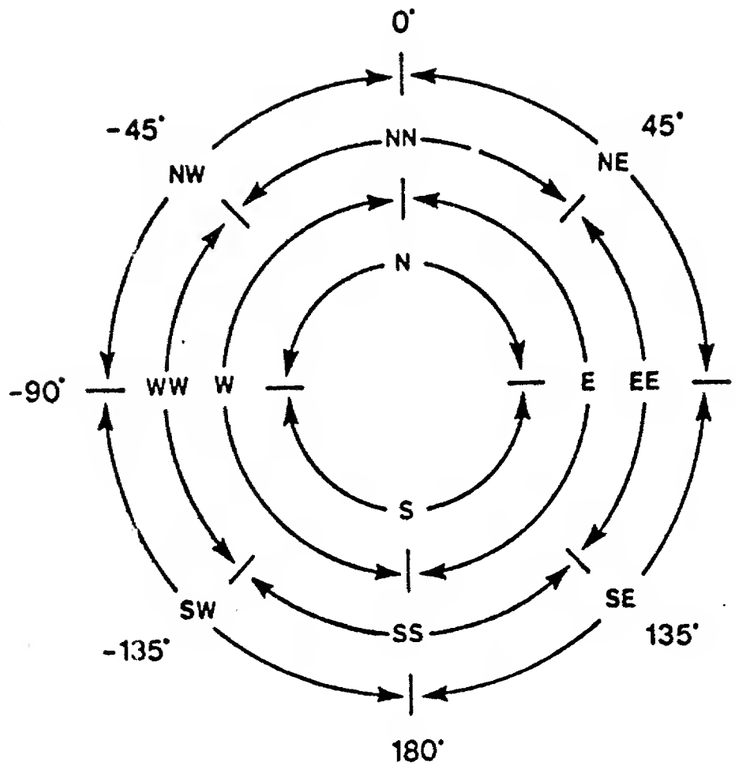
Having dealt with changes in the shape primitives of the 3-D model representation, we are left now with motion of the axes themselves--that is essentially, the problem of describing the movements of stick figures. There are three aspects to this: (i) changes in the length of an axis, which corresponds to changes in the overall size (*s* of the last section) of the attached shape primitive; (ii) motion of an overall model axis; and (iii) motion of one axis relative to another. For each of these cases, the critical point lies in discovering the natural coordinate frame in which to describe the motion.

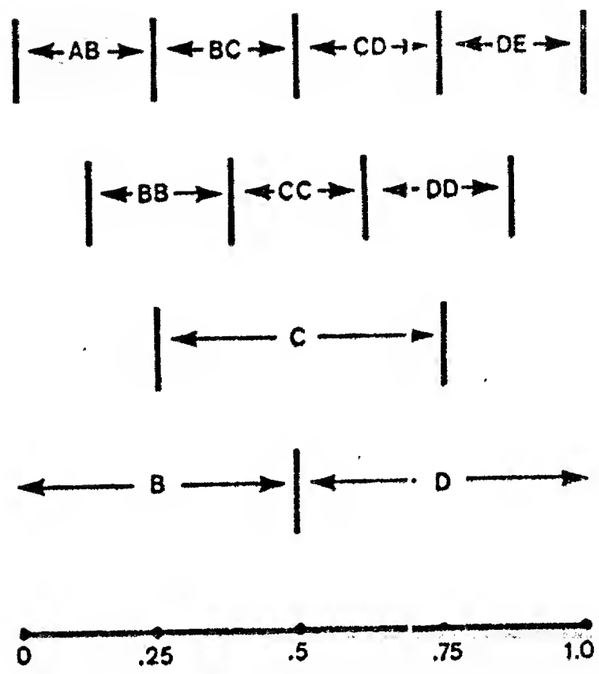
(i) Changing axis length

In the spirit of the previous section, the representation of changes in axis length *l* will be restricted to its time derivative *dl/dt* or *l*, written in a specification system that includes tolerances so that the accuracy of the parameter's value is also made explicit. Again, there is a question of whether and how to normalise the value of *l*; for a component axis within a 3-D model, the normalized values *l/l* are usually more useful than pure *l*.

(ii) Motion of an overall model axis

From a dynamical point of view, it is natural to decompose the description of a body's motion into two components, corresponding to the body's linear and angular momenta, because in the absence of net external forces and torques (a surprisingly common situation), these remain constant. Provided that the body retains constant mass and shape, linear momentum depends on the linear motion of the body's centre of mass, and angular momentum on the body's angular rotation. This can be measured either about the body's centre of mass, or about a stationary point on the body (for example).





A	S	p	r	θ	i	φ	s
model	torso	BC	AB	NN	NN	NN	CC
torso	head	DE	AB	NN	NN	NN	BB
torso	arm	DE	BB	EE	E	E	DD
torso	arm	DE	BB	WW	E	W	DD
torso	leg	AB	BB	EE	SS	NN	DE
torso	leg	AB	BB	WW	SS	NN	DE

Figure 2: From Marr & Nishihara, figure 5. The dispositions of the component axes in a 3-D model are described in terms of a set of angles and lengths called an adjunct relation. Angle and distance specifications in an adjunct relation must include tolerances so that the specificity of these parameters can be made explicit in the representation. One way to do this is shown in the upper diagrams which associate symbols with angular and linear ranges respectively. An example of adjunct relations for the human 3-D model in figure 1 using these symbols is shown in the lower table. *A* and *S* identify the two axes related by the adjunct relation specified on each row. If these mnemonic names were replaced by internal references to the corresponding 3-D models whenever they exist and left blank otherwise, this table would show essentially all the information carried by a 3-D model.

If we think of the model axis of a shape as roughly defining the distribution of the shape's mass in space, it becomes natural to describe the motion of such a body by splitting it into two components; (i) the overall motion of the model axis, and (ii) its rotation measured either about its centre or about a stationary point on it. For a walking man, the rotational component is zero (except as he turns a corner), but the translational component is not; for a child cartwheeling or rolling down a bank, both are non-zero; and for a pirouetting ballerina, only the rotational component is non-zero.

The representation of the motion of the centre of gravity can be carried out either in some external frame of reference--e.g., towards the viewer, towards the big tree--or in an object-centred frame, e.g., forwards, to the left. Representing the rotational component requires only two axes, the model axis and the axis of rotation, together with an indication of the inclination of the two axes (which will be roughly constant) and the angular velocity. For a pirouette or for the rolling child, the axis of rotation coincides with the shape's model axis, (i.e., the angle of inclination is zero). For a cartwheel, the inclination angle is 90^{o}.

Finally, when one end of a rotating shape is held still in some way, the resulting rotation is best related to the stationary point. For example, a blade of grass, a stem of wheat, or a swaying man are all in some sense affixed to the ground; a bat or a monkey hanging from a tree are also essentially pivoted. In such cases, the swaying or falling motions simply require specification of a direction and rough speed; and once again this can be carried out either in an object-centred or viewer-centred frame, or even in the frame defined by some external object, whichever is most useful for the purpose at hand. The important point is to decompose the motion by taking advantage of the pivot, and then to describe it in the most advantageous coordinate frame.

(iii) Motion of one axis connected to another

The final case, and the most important one for the purposes of this article, concerns the situation when an axis is pivoted about a point along it, for example at one end. Here it is visually natural to represent the motion in a coordinate frame centred on or near the pivot point, so that for instance the motions involved in swinging an arm would be referred to a frame centred on the shoulder. It is perhaps worth noting that, in contrast with case (ii) discussed above of the motion of an overall body, the visual simplicity of such a description is not matched by any underlying dynamical simplicity. If the shoulder is itself accelerating, the effects on the arm and hand may be complex.

In the 3-D model representation, the position and size of one axis is represented relative to another by an adjunct relation *(p, r, θ, i, φ,* s). The numbers *(p, r, θ)* represent in cylindrical polar coordinates the connecting point of the axis; for an arm, *(p, r, θ)* would define essentially the location of the shoulder joint relative to the torso axis (see figure 1). The remaining numbers *(i, φ, s)* specify the size *s* of the arm relative to the torso, and its orientation in spherical polar coordinates. The angle *i*, called the inclination, is the angle between the torso axis and the arm axis; thus it is 0^{o} if the arm lies parallel to the torso, and 90^{o} if the arm is held straight out horizontally. The other angle φ measures the declination of the arm. For a horizontally held arm, it would be 0^{o} if the arm pointed straight ahead, 90^{o} if it pointed to the right, 180^{o} if it pointed straight behind, and so forth. All the angles and sizes are presumed to be written in a scheme that defines both a value and a tolerance (like that of figure 2).

The problem of describing motions of a pivoted axis, such as an arm or leg axis, is more difficult than that for a whole human model axis because complex forces may be transmitted across the junction. This means that in principle, arbitrary motions can occur. Instantaneous motions can, however, be captured by measuring di/dt and $d\varphi/dt$, and as we shall see later, these alone go far towards being an adequate representation of motion for the recognition of common types of movements and gestures, like walking, climbing, saluting and

so forth.

Rotations, however, are not easily captured in a representation that makes explicit only di/dt and $d\varphi/dt$, unless the rotation is such that either i or φ is constant. For this reason, we introduce one more primitive into the representation of the motion of a pivoted axis, namely a primitive for the circular rotation of the pivoted axis about an axis of rotation passing through the pivot point. Thus if one extends one's arm horizontally out to the side and makes circular movements, this would be represented as a rotation about the axis $i = 90^{o}$, $\varphi = 90^{o}$. To specify the rotation completely, two other numbers are needed: the angle of inclination of the rotating shape axis to the axis of rotation (this might be 10^{o} for the small circular movement described above), and the angular speed of the rotation. The axis of rotation can be specified by means of an adjunct relation, and the two additional parameters can be specified, with tolerances, in the usual way.

An Instance of the Representation

In order to illustrate the use of this representation in describing the instantaneous motion of a shape, we show in figure 3 the instantaneous description of a walking human shape. At the top level, of the model axis for the whole human shape, the speed and direction of the walk are captured in object-centred coordinates. The main aspects of walking, the contrapuntal swinging of the arms and legs, are captured in the specifications of the movements of the human component axes, as shown. Further detail is provided by subsidiary LEG 3-D models, which show that one of the legs is bending at the knee, whilst the other is not. The arms and legs in the drawing of the figure have perforce had to be given a position in order to be shown, but this information is not included in the motion model given in the figure. To add this information, however, one simply adds a shape 3-D model to the description.

Characteristics of this Representation of Instantaneous Movement

At this point, it is perhaps worth examining the representation defined in the previous sections in terms of Marr and Nishihara's three criteria.

I. Accessibility

The recovery in real time of instantaneous three-dimensional angular velocities from two dimensional images is one of the things which, while possible in theory, is in practise difficult to do accurately. On the hand, accuracy is probably not very important here, for as we shall see in the next part of this article, few if any common movements require great precision, and for many the sign is enough. In general, provided one is given a little time, it is not unreasonable to expect that one can recover the rate of change of angles like i and φ that are already explicit in a 3-D model.

Description of rotations are always simplest when referred to a coordinate frame based on the axis of rotation. For many common movements, the axis of rotation coincides with the axis about which one of the 3-D model angles is measured. For example, while swinging the arm back and forth during walking, φ remains constant and only i changes, so the appropriate angle to use here is in fact i. But it is perfectly possible to perform motions in which this is not the case; for example, an arm motion like making small circles round a direction up and to the right involves changes in both i and φ. At any moment, the motion is a rotation about some

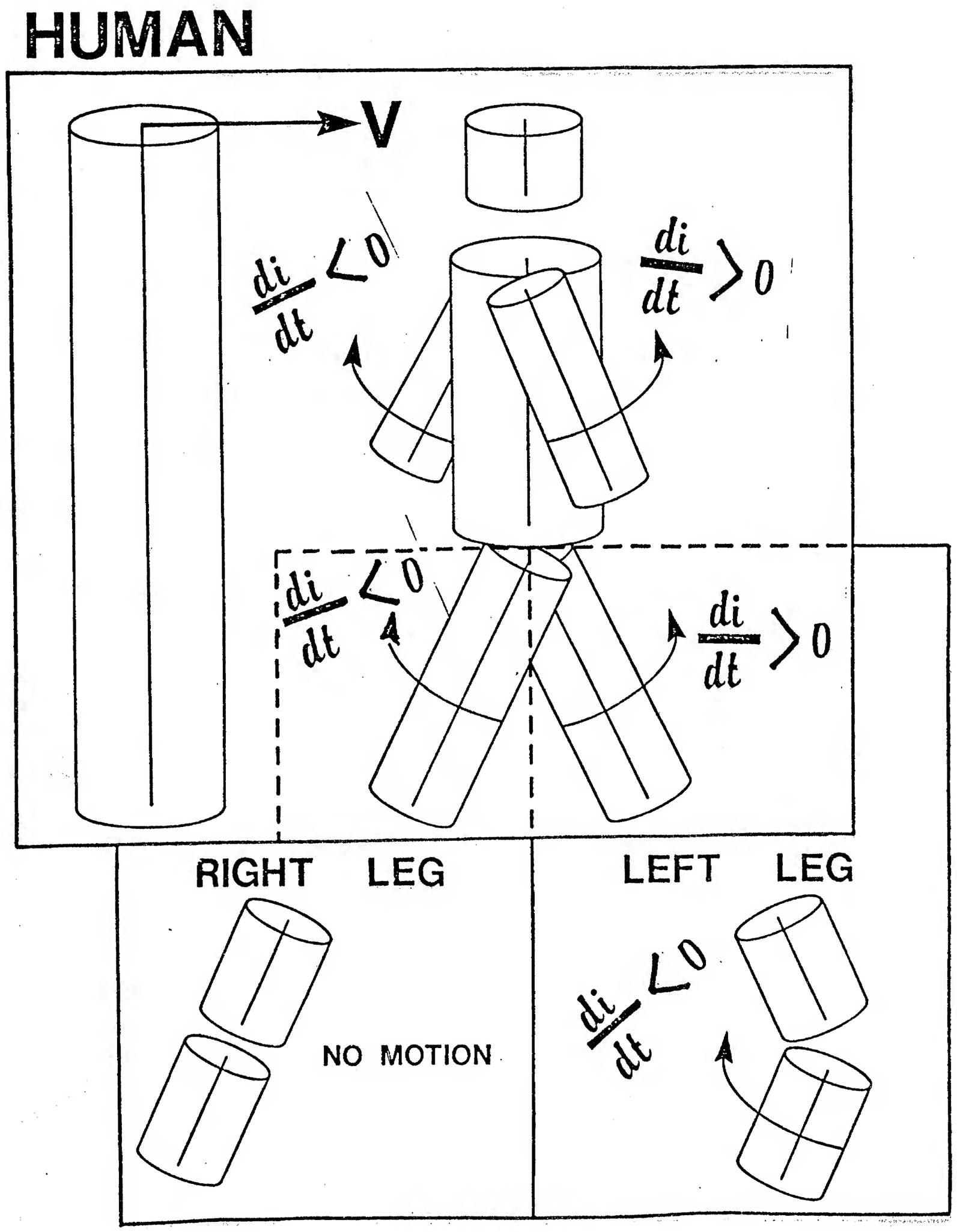


Figure 3. This figure illustrates the proposed representation of the instantaneous movement of a shape, here of a walking man. It makes explicit (a) the overall velocity (speed and direction) of the walk, i.e., that it is forwards with speed v; (b) the arms and legs and swinging in the normal walking pattern (signified by whether di/dt > or < 0); (c) one of the legs (the left) has a subsidiary motion consisting of bending the knee; (d) the right (load-bearing) leg does not have this motion.

axis, of course, but in practise it is likely to remain difficult to recover a precise enough estimate of di/dt and $d\varphi/dt$ to allow a reasonable guess at where this axis is, unless the motion continues for a considerable time.

If the motion is a rotation, and if it is prolonged, then several methods will be capable of recovering it. Provided the axis is associated with a rigid non-planar surface, Ullman's structure-from-motion theorem shows that it can be recovered quite quickly. If only a stick is moving, so that not enough non-coplanar points are available, the rotation may still be recognizable provided that it passes through more than a cycle.

2. Scope and uniqueness

Because it is so closely tied to the 3-D model representation, the scope of the shapes to which the representation applies, and the uniqueness with which they are represented, are the same as for 3-D models. For the motions, the scope of the representation is limited to velocities; no acceleration or higher terms are included. And in addition to this, arbitrary changes to the shape primitives are not easily captured by the representation.

Finally, although the representation is defined formally only for instantaneous movements, we shall in the next section be using one description to cover the state of affairs over an extended period of time. For example, the descriptor ($\varphi = 0$, $di/dt > 0$) of figure 3 remains true over the whole of one arm-swing during walking, thus allowing a true if somewhat coarse representation of the state of affairs during the whole of that time. This type of idea works well when motion segments are separated by stationary moments, as they are during walking. It is, however, difficult to apply to rotations, which are a form of motion that can last for an arbitrary length of time without any stationary moments. If one swings one's arm round and round, the motion never ceases. It is for this reason that the representation of a motion as a rotation will be of particular importance to us later in the article, and it is why we have introduced it as a primitive here. The scope of our representation of rotational movement is however, quite restricted; in specifying only an axis, an inclination and an angular speed of rotation, we have essentially limited ourselves to constant circular rotations. Varying angular speeds or elliptical orbits do not fall within the scope of the simple representations we have defined. Although one can conceive of extending it to these cases, we shall not need to for what follows.

3. Stability and sensitivity

The third aspect of this representation is also heavily dependent on the 3-D model representation. Just as for shapes, the motions here are described using a modular, hierarchical organization. Thus in figure 3, the overall motions of the limbs during walking are captured in a stable manner at the level of the HUMAN module, whereas the fact that one leg bends at the knee during the step is made explicit in the LEG module, at a more detailed level of shape description. This more sensitive information will in general be more difficult to recover in real time.

Representing Movements that are Extended in Time

Except for a brief excursion into the subject of rotational movements, our discussion has hitherto been confined to the representation of instantaneous motion, of the kind illustrated in figure 3. The essence of motion is however that it causes changes in time. The description of a shape in motion can thus rapidly become outdated. For instance, in the case of the walking example of figure 3, the particular description of instantaneous motion given there will remain valid for one half of the walking cycle, but for the other half, the "positives" and "negatives" would have to be reversed. A description that included a specification of the instantaneous positions of the limbs would cease to be correct even sooner. There is of course a trade-off

involved: the coarser the tolerances in the representation of numerical values associated with the primitives, the longer the description will remain valid. For example, if *di/dt* were expressed in a system with smaller tolerances, say (-fast, -medium, -slow, zero, +slow, +medium, +fast) instead of just (negative, zero, positive), then the description would be valid for a shorter time, although during that time, the information made explicit would be more accurate.

Nevertheless, in dealing with an extended movement, one will at some stage have to confront the problem of assembling several such descriptions, corresponding to the different parts of the movement as they are played out in time. The critical question is, of course, how to do this? What is a "part" of a movement? Where does one piece of a movement stop and another begin? *How* and *why* may one "segment" a movement?

This question is reminiscent of one that commanded considerable attention in earlier vision studies, namely, how should one "segment" an image into meaningful objects, and what in any case was an object? It is clear that there is an underlying truth involved here, namely that in the physical world, matter is cohesive and arranged into separate bounded pieces. But this is hard to apply directly to images; is a man on horseback an object, for example? Is a head one, or a nose? These issues were resolved by the introduction of the 2-1/2-D sketch and 3-D model representations, one of which reflects the physical realities of the visible surfaces, whilst the other allows one to assign a separate description to a piece of a shape whenever it is descriptively convenient to do so (see Marr, 1981).

In the temporal case we have a similar situation, but the objective criteria for separation are unfortunately not as strong as those supporting the decomposition of the material world into objects. For this reason, semantic and interpretive aspects can be expected to intrude more deeply into the analysis than they did for the representations of static shapes.

The basic fact about the motion of articulated shapes that we need here is the observation that during a given movement, motion often ceases. The arms swing during walking movements, but twice each cycle they are instantaneously at rest, relative to the body. Of course, some motions never include stationary moments, and this is why rotations give so much trouble. Often, however, such motions involve repetitions--they constantly return to the initial configurations--and this can be used to decompose the motion into suitable descriptive units. For overall motions of whole bodies like a walking man, stationary points are rare, and we can use instead discontinuities in the speed or direction of movement. Arbitrary motions, however, pose just as difficult a representational problem as do arbitrary shapes.

The SMS Representation of Movement

The fact that movements often include pauses, i.e., moments when parts of a shape are either absolutely or relatively at rest, allows us to define a natural or canonical decomposition of a movement into a sequence of *motion segments* and intervening (*static*) *states*. Thus, for walking, there are two "states" and two "motions"; the two states are those when the arms and legs are instantaneously at rest relative to the torso; and the two motions are those occurring in figure 3,and its counterpart obtained by interchanging "positive" and "negative". Notice that the arms and one of the legs are not really at rest (relative to the ground) in the two extremal states, because of the forwards motion of the walk itself. But they are at rest in the (object-centred) coordinate frame centred on the torso. We therefore define the *state-motion-state* decomposition of the movement of a *single 3-D model* in the following way.

Definition. Let M be the model axis of a 3-D model. Let P be its principal axis (local coordinate axis) and let $C_1, C_2 ... C_n$ be its component axes (see e.g., the human 3-D model of figure 1). Then (i) if M is at

rest in the coordinate frame to which it is currently referred, we say that the 3-D model's shape is in an *overall rest-state*; (ii) if axis C_i is at rest relative to P, the shape is in a *local rest-state*. (iii) If all the component axes are at rest relative to P, we say that the model is *static*.

Thus, for example, if a walking man starts running on the spot, his human 3-D model representation (assuming it is accurately computed) will be in an overall rest-state. If he stands still on an escalator, his 3-D model will be static, but he will not be in an overall rest-state (relative to an external observer). If he stands still but waves his arms around, his 3-D model will be in local rest-state defined by the legs.

As we saw in figure 1, the full description of the shape of a human includes many 3-D models, arranged in a hierarchy. We can capture the essence of even quite complex movements in a similar way. For example, take the case of the gentle throwing motion usually called LOB; in one version of this, one stands still, swings one's arm forwards keeping the elbow joint straight, and lets go the ball by opening one's hand at the end of the swing. In this movement, the overall HUMAN 3-D model would be at rest; the arm axis C_2 (say) will be moving relative to the human's principal axis, so the ARM 3-D model is not in an overall rest-state. Since the elbow does not bend, however, the ARM 3-D model itself is static. The FORE-ARM model is however not always static, because the HAND component axis moves as the ball is let go.

This sequence is illustrated in figure 4. To begin with (a) the man is static, about to begin the movement. In (b) the arm swings, but there is no motion at the lower levels of description. In (c) motion appears down in the FOREARM module, the ARM module is still static, and the HUMAN module displays the same motion as before. In fact, we have also included some positional information in the HUMAN module, as well as its motion, for the arm is shown further on in its rotation than in (b), roughly specifying the position at which the hand joint in the FOREARM module starts to move. Next in (d), the fingers move in the HAND module, as the thrower releases the projectile. Finally, in (e) the movement ends and all 3-D models become static again.

Two points emerge from this example. Firstly, the action LOB involves essentially just one arm; it matters little whether the lobber is simultaneously moving his head, legs, or other arm. It is initiated and terminated at the largest scale by local rest-states in the HUMAN 3-D model. These are overall rest states for the ARM 3-D model. Secondly, although the overall description of LOB consists of an arm swing, a much more complete description may be obtained by examining the states of the subsidiary 3-D models during the motion. These are the ARM, FOREARM and HAND 3-D models, and the important subsidiary rest-states occur just as the arm starts to rotate, and the fingers rotate to ungrasp the projectile. In all cases, for the ARM, FOREARM and HAND 3-D models, the critical moments for segmenting the movement occur when the model changes from an overall rest-state (i.e., a local rest-state from the point of view of the model's overall coordinate frame) to a moving state. The rest-state itself may be prolonged (as in lob) or instantaneous (between the swings in a walk); but in either case, it is the *change* that marks the temporal boundary and is incidentally often easy to detect visually, e.g., through the mechanisms of directional selectivity (see Marr & Ullman, 1980).

Using this criterion for the decomposition of a movement allows us to define a canonical, hierarchical representation of the movements of shapes that admit of a 3-D model representation. We set up the definition as follows:

Definition (i). Suppose that we are given the 3-D model representation of a moving shape, and let M be a particular 3-D model within it (e.g. the ARM model for a human). Then a *motion segment* for M consists of the interval between two adjacent overall rest-states of M.

It is clear that any extended movement of M can be split up into motion segments, and that for many

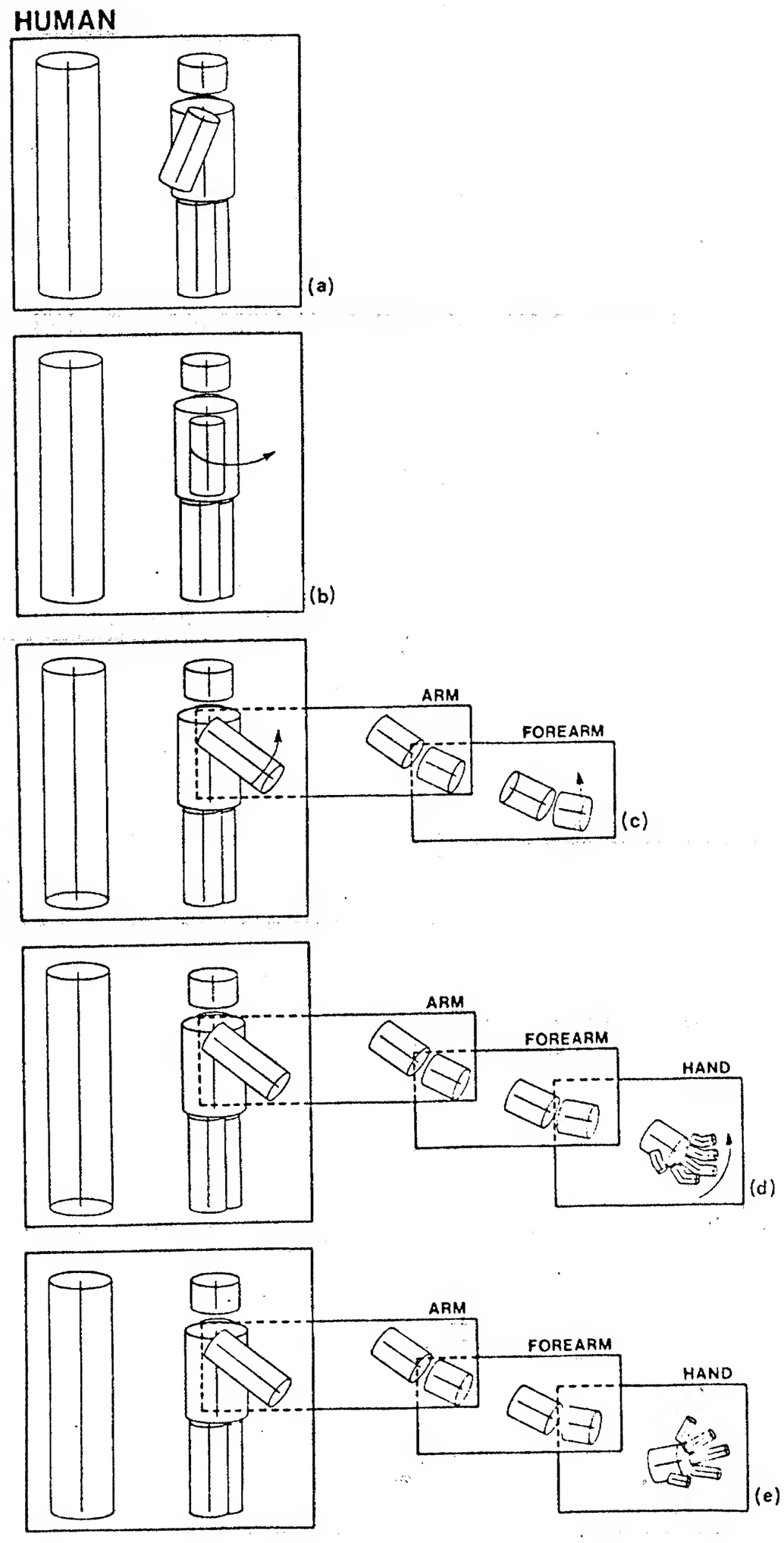


Figure 4. The sequence of movements that constitute the overall movement, LOB. To begin with, (a) the man is static, about to begin the movement. In (b) the arm swings, but there is no motion at the lower levels of description. In (c) motion appears down in the FOREARM module, the ARM module is still static, and the HUMAN module displays the same motion as before. In fact, we have also included some positional information in the HUMAN module, as well as its motion, for the arm is shown further on in its rotation than in (b), roughly specifying the position at which the hand joint in the FOREARM module starts to move. Next, in (d) the fingers move in the HAND module, as the thrower releases the projectile. Finally, in (e) the movement ends and all 3-D models become static again.

movements this is a useful decomposition. Decomposing a movement in this way allows us to define a canonical representation of it. Each motion segment of the movement is described using the instantaneous motion representation defined in the first part of this article. Each intervening state is described using the 3-D model representation for shape. Finally, these two types of description are combined into a sequence, state-motion-state-etc. thus creating a representation of the whole movement.

Definition (ii). This representation of the (extended) movement of a 3-D model is called the *state-motion-state* (SMS) representation of the movement of a 3-D model.

This gives us the representation of movement at one level in the 3-D model hierarchy, and it only remains for us to include the facility for the more detailed descriptions of motions that are allowed by more detailed descriptions of shape, as we did for examples in figures 4 (c) and (d).

To do this, it is necessary only to allow motions of lower, more detailed 3-D models to be added to the overall representation of the movement. For example, at some point, we needed to add the ungrasping of the hand to the rotation of the whole arm during the action LOB. The point at which this is added is, of course, defined by the motion segment of the HAND 3-D model; but the important point here is that at the moment the HAND begins to move, higher level 3-D models (like the ARM in the HUMAN frame) may already have been in motion for some time. We therefore include in the representation not only a specification of the HAND 3-D model's initial state (as specified in the FOREARM 3-D model) but also a description of (roughly) where the ARM has got to at that moment (as we did in figure 4 (d)).

We are now in a position to give the full definition of the SMS representation, as applied to the whole of a 3-D model shape representation.

Definition (iii). As in definition (i), suppose that M is a particular 3-D model included in the 3-D model representation of a moving shape. For M, and for each 3-D model below M in the shape hierarchy, we create the SMS representation of the movement. In addition for each model below M, for each state description in the SMS representation, we include the current shapes and motions of its superior 3-D models, up to and including M.

Rotational Movements

We saw one rough example of the SMS type of representation in figure 4. The decomposition upon which the scheme rests depends however very much on the occurrence of stationary states during the movement. As we saw earlier, rotational movements pose special problems for such a scheme, since they involve no stationary states. Provided rotational movements can be recognized as such, we can simply add them to the representation, considering rotation primitives as a special type of motion segment description, and treating them in the same way.

Translational Movements

Finally, we saw earlier that dynamical arguments suggest forming a separate description for translational movements of a whole shape. Segmenting such movements only at stationary points would often provide too impoverished a description of the motion, so one can add to them points where the direction or speed of the motion change discontinuously. For a walking man, the segmentation points would then include where he starts and stops and changes direction. Such a decomposition would probably suffice for the requirements of

everyday life, but it falls far short of a comprehensive representation of arbitrary trajectories.

In this way, we arrive at a primitive but fairly comprehensive way of representing shapes in motion. We call it the *SMS moving shape representation*, and in the next section, we give some examples.

Some Examples

In order to show how the SMS moving shape representation may be used to describe movements, we analyze four common examples, walking, saluting, kicking, and a simple kind of throwing.

The representation of walking we have separated into two figures. The first, figure 5, depicts part of the SMS representation for one half of the walking cycle. The second figure 6 portrays more clearly the component of the SMS moving shape representation that deals with one leg during a step. The LEG 3-D model (a) consists of a model axis for the whole leg, which is identified with the LEG component axis in the higher-level HUMAN model; and two component axes, for the upper and lower parts of the limb. In the HUMAN model, the SMS representation of the leg motion is illustrated in (b), consisting of just one motion segment. The angle *i* here refers to the inclination to the torso axis (shown dotted), and the angles have been written as points of the compass (as in figure 2) to indicate that they are only approximate. The positions of the sticks are shown in the figure in viewer-centred coordinates, for convenience a slightly improper depiction of the representation. The motion of the component axes is referred to the local leg coordinate system, (shown dotted in (d)). The knee is initially straight, i=SS, the angle *i* being here the angle between the lower segment of the limb and the local axis. The knee bends ($di/dt < 0$) and becomes stationary (i=WW), then the sequence reverses. In this simplified model, the last motion sequence ($di/dt > 0$) occurs during the final end state of the model axis S, as depicted.

Our next example is of a salute, and in figure 7, three levels of representation are shown. At the coarsest level, the level of the human model axis, the HUMAN 3-D model is in an overall rest state. At the second level associated with the component axes of the HUMAN 3-D model the arm moves ($di/dt > 0$), the angle *i* here refers to the inclination of the arm to the torso axis. The critical aspect of the salute is captured in the ARM 3-D model. The *i* here is between the upper arm and the lower arm and $di/dt > 0$.

In the next examples, we wish to introduce a slightly different set of ideas, namely the visual precusor of the notion that an action can involve an object. "Kick" is a simple example of an SMS representation involving a projectile. The HUMAN 3-D model representation is in an overall rest state, the movement is captured in the LEG 3-D model. Figure 8 shows the leg and the projectile, both at rest. The angle i = SE refers to the initial inclination of the leg whithin the HUMAN coordinate frame; then the leg moves towards the projectile (di/dt >0) After the projectile starts to move in the LEG frame, the leg stops (i = SW). Finally we see the projectile moving relative to the HUMAN model axis.

In our analysis of LOB (see figure 4) our description was confined to the movements made by the thrower (e.g., the acts of swinging the arm and ungrasping the projectile). A critical aspect of the action is however that the projectile becomes detached from the thrower, and this is captured at the coarsest level in figure 9. This figure shows the top-level representation of LOB from figure 4, with one addition: the standard HUMAN-ARM 3-D model has been replaced by a new 3-D model, consisting of one stick for the arm, and another for the projectile at the end of the arm. This 3-D model is static until the "ungrasp" action in figure 4, at which point the projectile axis begins to move in the arm frame. The motion here consists of a change in *r* (of the adjunct relation), and not (as usual) of *i* or φ. Finally, we have shown the projectile moving relative to the human's model axis, thus making explicit the direction in which the projectile has been thrown.

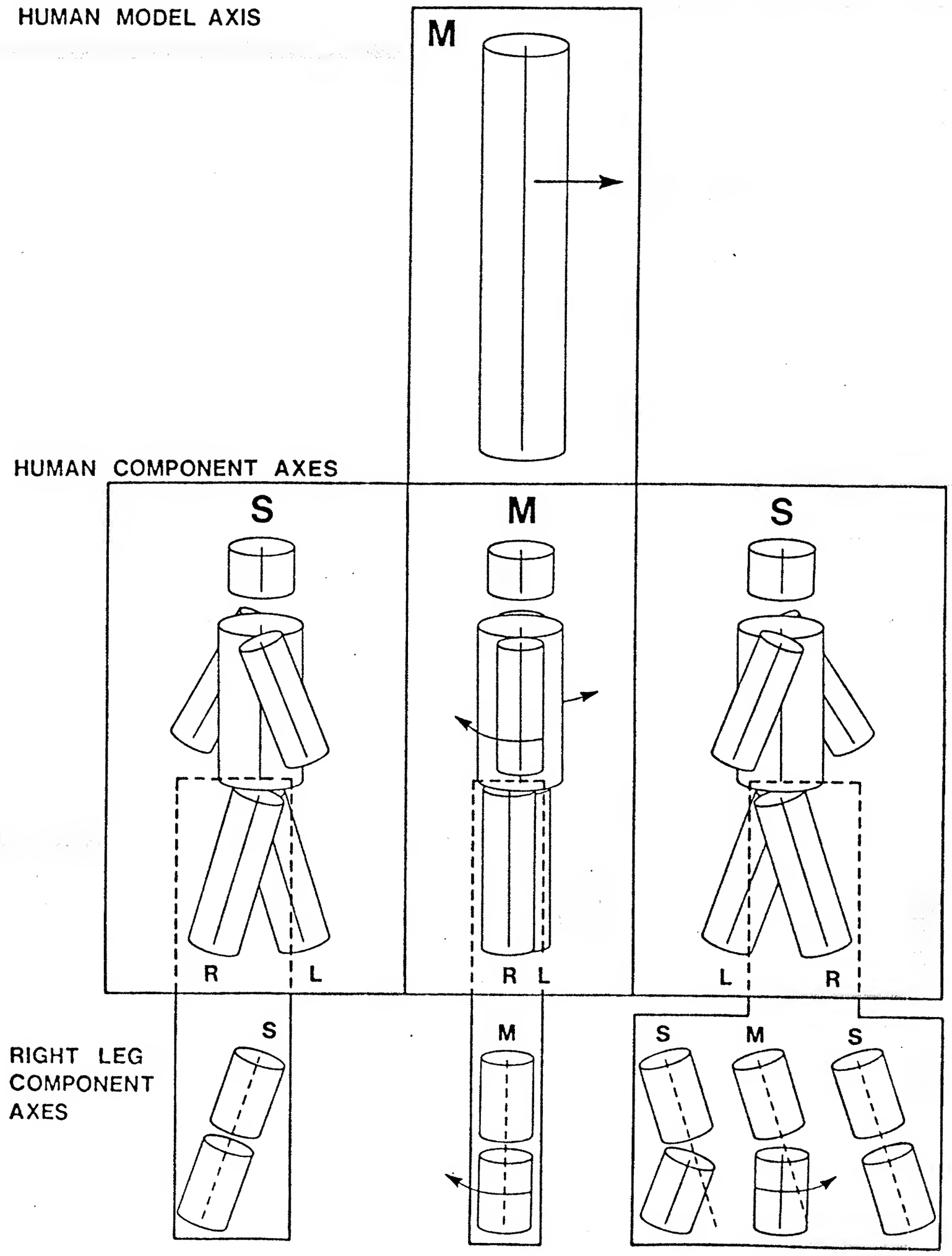


Figure 5. This and the next figures give visual depictions of the SMS representation of various types of movement. This figure the walking sequence (S denotes stationary states of the 3-D model, and M ones in which motion occurs) at a level at which the shape is only coarsely described. A detailed description appears in the text. Three levels of representation are shown: (i) the overall motion of the walk, captured at the level of the human model axis, (ii) the swinging of the arms and legs, captured in the motions associated with the component axes of the HUMAN 3-D model, and (iii) the motion of the knees of the non-load-bearing leg. By attaching additional FORELIMB 3-D models to the representation, the motions of the feet during a step can be represented in a similar way.

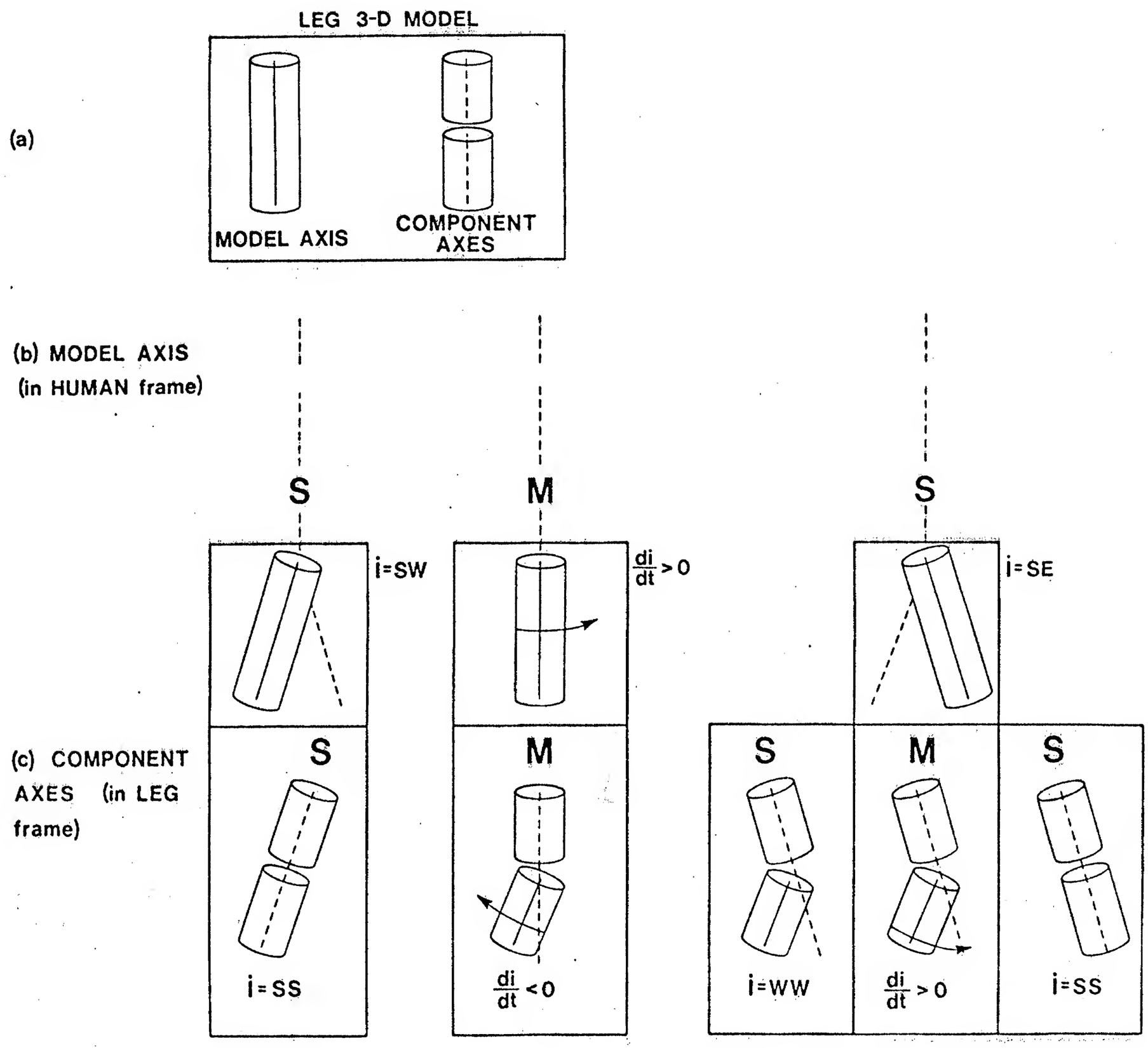


Figure 6. A clarification of Figure 5 at the more detailed level of description of the walking shape. Just as an arbitrary level of detail about shape can be included in a 3-D model description, so can an arbitrary level of detail about a movement be included in the SMS representation.

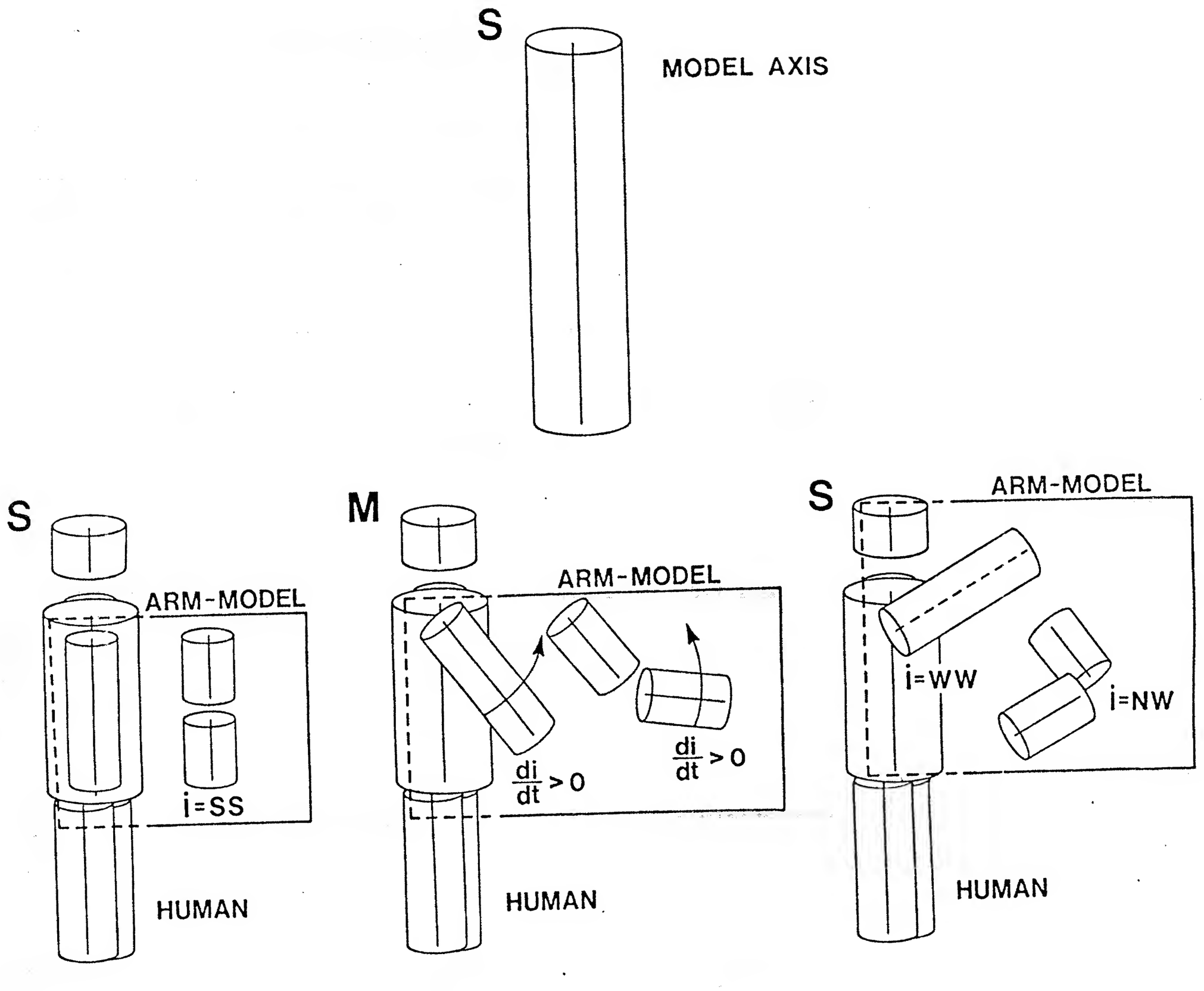


Figure 7. A visual depiction of the SMS representation of a salute. The motion includes two motions, at the shoulders and elbow.

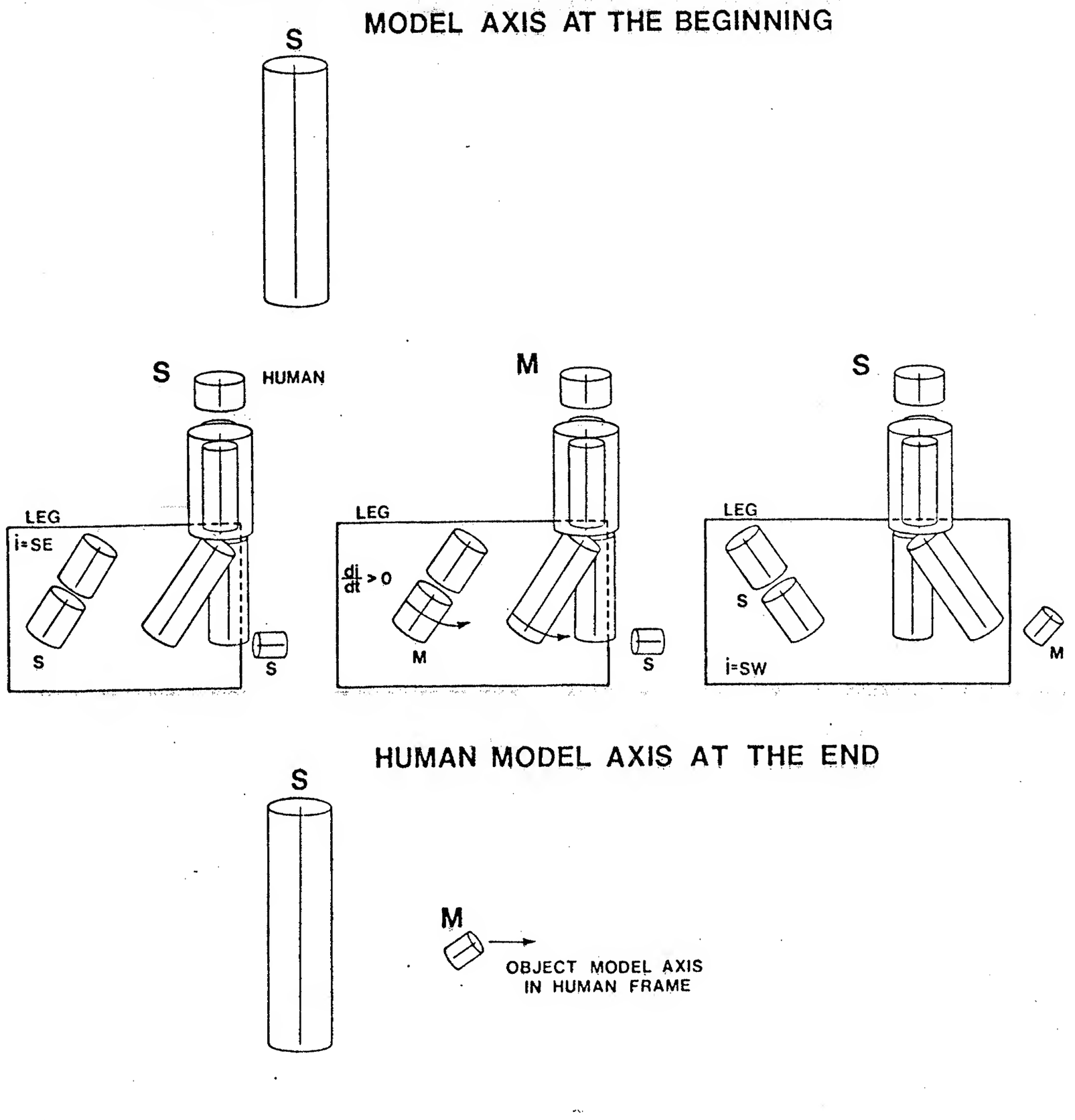
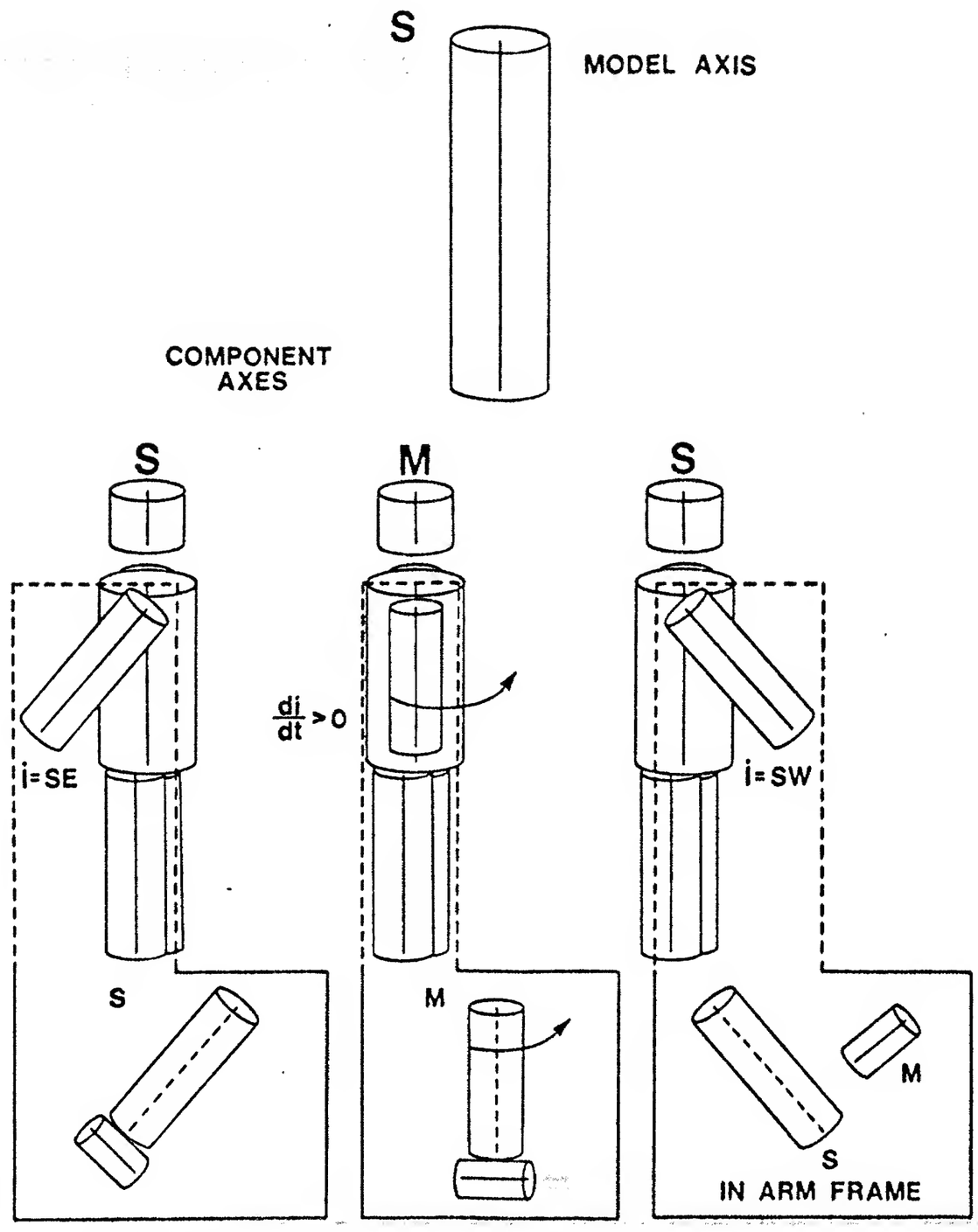


Figure 8. Many types of movement correspond to actions that involve objects. Here we depict a kick. Initially and finally the human shape is overall at rest (S) but in the final state, the projectile is moving (M).



FINALLY

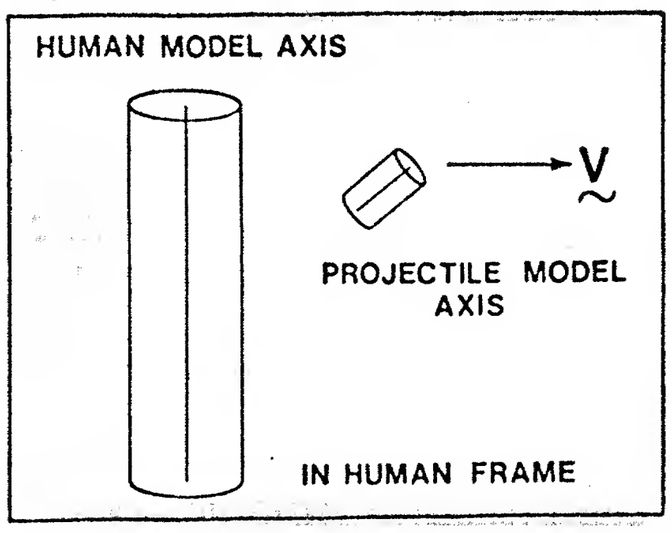


Figure 9. This shows the lob action again, but the detail of Figure 4 has been omitted. Instead, the object being lobbed is included in the description, showing as in Figure 8 how the SMS representation leads naturally to questions about the relation between movements of 3-D shapes, and the notions of object and action.

At the rather coarse level of description portrayed here, the "ungrasping" action of releasing the projectile is not represented in any detail. In the full SMS representation, the ungrasp act would form an explicit motion segment associated with the HAND 3-D model, along the lines depicted in figure 4(d). At this top level, however, only the grosser aspects of the motion are made explicit. The detailed description of "ungrasp" here is probably hard to derive visually from just one occurrence, but the coarse one of Figure 9 is often quite easy, because the visual system contains many quite simple mechanisms for noticing the motion of an isolated projectile.

Deriving the SMS Representation from Images

The construction of the SMS description of a movement involves (i) the construction of 3-D models for the moving shapes, (ii) splitting the movement into motion segments, and (iii) assembling the SMS representations for each segment, to the degree of detail required by the viewer. In their discussion of the derivation of a 3-D model for a static shape, Marr and Nishihara showed how interaction between information from the image and a stored catalogue of 3-D models can often help the construction of a more accurate description of the viewed shape (see their figure 9). In order to make this type of derivation possible, the 3-D model catalogue needs to be indexed in several ways, for example from general to particular shapes i.e., cylinder biped model man model (see their figure 8).

The end result is to derive a description of the currently viewed shape using the best matching models in the catalogue, and this is how they viewed the recognition-derivation process. Similar issues arise here, in dealing with the representation of movements in time as well as with the underlying representation of the shapes that are moving. Table 1 lists some of the movements of a human shape that have straightforward descriptions in the SMS representation along the lines of figures 5 - 9, and which one would expect to find in one's catalogue of movements. In this particular table, the movements have been organized roughly by the piece of body involved, and this is probably one useful way of indexing it.

There are however several other ways that are probably important. One is by a movement's SMS components. For example, the M module of a HUMAN 3-D model in mid stride is often enough to diagnose a WALK; and some S descriptions are also quite characteristic of a particular movement, for example the beginning of a tennis serve, or the end of a LOB. And just as for static shape, movements can be organized along the coarse-to-detailed axis, in two senses; firstly as induced by the shape hierarchy, from the movements of the coarse shape to movements of the detailed components. For this, one requires indexing by specificity of shape, like Marr and Nishihara's figure 8; and one needs backwards indexing, so that a characteristic motion of a small component of the shape can give useful information about the overall movement. For shapes, this type of indexing (which Marr and Nishihara called *parent* indexing) allowed recovery of the horse 3-D model from the model for a horse's face. For movements, such a thing would allow for example a particular type of motion of a hand to lead to the recognition of a particular type of THROW, perhaps that involved in throwing a coit or a frisbee.

The second aspect of the coarse-to-detailed organization of movements would be the kind that mirrors the specialization of a rough kind of LOB for a particular task, for example, to the particular type of LOB involved in bowling. Here it is not the shape components themselves than become more specialized, it is the timings, angles, and velocities that become more precise. Thus it will be important not only to be able to make explicit the tolerances involved in the representations of these numerical quantities, but also to be able to index into the catalogue along these dimensions.

SOME COMMON MOVEMENTS

TURN-ROUND FALL-OVER ROLL SOMERSAULT CARTWHEEL PIROUETTE SWAY	whole body	PRESS-UP SALUTE BECKON SEMAPHORE WAVE THROW	arms
NOD - YES - NO - QUESTIONING	head	STEP WALK RUN MARCH LIMP HOP JUMP	mainly legs
BEND-OVER SIT TOUCH-TOES ROW	mainly torso	CRAWL CLIMB BUTTERFLY-STROKE CRAWL (swim) BREAST-STROKE	arms and legs

Table 1. Some Common Movements.

DISCUSSION

The basic question which Marr and Nishihara asked was, how may static shapes be represented for the purposes of visual recognition? They restricted their inquiry to the shape of objects (i.e., disjoint pieces of solid matter rather than fluids), and their representation was based on decomposing the shape of an object into components associated with the natural axes of the shape. Some shapes, like animal shapes, are easy to represent in this way, and some, like a crumpled newspaper, are not.

The representation that they proposed, the 3-D model representation, provided us with a reasonable starting point for our enquiry into the problems posed by moving shapes. The additional dimension introduced here is time, and our basic problem was to find a canonical way of segmenting into components the continuous stream of data provided by a moving shape. The underlying difficulty of the problem arises because in real life, there are few criteria for establishing demarkations in time as objective as those that separate objects or define axes in space.

The solution we proposed to the problem rested on the idea of a motion segment, which is a piece of movement bounded by stationary states. As we saw, the notion of a stationary state depends on the coordinate frame one chooses, and this in turn depends on the particular shape representation being used. An arbitrary motion need not have any stationary states, or they may occur only seldom. Such motions would be hard to represent accurately using our scheme, just as a crumpled newspaper is hard to represent in 3-D models. Some types of unsegmented motions, particularly repetitive ones, are however both common and important, so for them we suggested special additional steps like including rotational motion as a primitive, and segmenting at discontinuities in velocity. Using these and the stationary criteria, we were able to define segmentation points for a considerable range of movements.

The representation itself consists of combining representations of the motion segments and their intervening states into a sequence. The result is a representation that combines the advantages of the 3-D model representation with those arising from variable precision in the description of the motion segments themselves. Finer detail in the representation can be achieved by including motion segments of subcomponents of the shape within the single state or motion descriptions of the larger shape components.

One can conceive of additional criteria for segmenting a movement, some purely visual, like sudden changes in acceleration, and some more cognitive, relating to the meaning of the action, cause and effect, and so forth. But our main point is that the simple, objective segmentation criteria provided by stationary states is sufficiently powerful to provide the basis for representing and discriminating amongst a wide range of common movements. Furthermore, in the light of the recent progress in understanding visual information processing, it is quite reasonable to suppose that the SMS moving shape representation can be visually derived.

REFERENCES

Hollerbach, J. M. 1975 Hierarchical shape description of objects by selection and modification of prototypes. *M.I.T. A.I. Lab. Technical Report 346.*

Longuet-Higgins, H. C. & Prazdny, K. 1980 The interpretation of a moving retinal image. *Proc. R. Soc. Lond. B* (in the press).

Marr, D. 1977 Analysis of occluding contour. *Proc. R. Soc. Lond. B. 197,* 441-475.

Marr, D. 1981 *Vision.* San Francisco: W. H. Freeman & Sons.

Marr, D. & Nishihara, H. K. 1978 Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B. 200,* 269:294.

Marr, D. & Ullman S. 1980 Directional selectivity and its use in early visual processing. *Proc. R. Soc. Lond. B* (in the press).

Sutherland, N. S. 1979 The representation of three dimensional objects. *Nature 278,* 395-398.

Ullman, S. 1979 *The Interpretation of Visual Motion.* Cambridge, Mass: M.I.T. Press.